Generation of Arm-gesture and Facial Expression for Intelligent Avatar Communications on the Internet

Sang-Woon Kim †, Young-Who Lee † and Yoshinao Aoki ‡
† Div. of Computer Science and Engineering, Myongji University, Korea
Email: {kimsw,yongwho}@mju.ac.kr
‡ Graduate School of Engineering, Hokkaido University, Japan
Email: aoki@media.eng.hokudai.ac.jp

Abstract

Recently the sign-language communication systems between avatars of different languages have been investigated as a means of overcoming the linguistic barrier. In the systems, an intelligent communication method has been employed, where sets of the animation parameters such as the joint angles of the gesture were transmitted instead of sending the entire-real motion pictures. However, the communication has been done based on the gesture only without considering the facial expression. In this paper we propose an approach to the communication based on the facial expression as well as the arm-gesture, and generating them on various avatar models. To extract the parameters to be transmitted, three kinds of key-frame editors are employed using techniques of inverse kinematics and partial differential equations. In generating facial expression especially, the movements of the cheeks and the jaws as well as other facial components are also utilized. The preliminary results show a possibility that the method could be used as a useful means for avatar communications between different languages on the Internet cyberspace.

1 Introduction

Internet cyberspace has become a place of connecting millions of people around the world based on multimedia such as text, sound and animation. To enter cyberspace and become a member of a new community, a virtual persona for self-representation, the so - called avatar, is needed [1]. When an avatar is navigating the space, he or she could meet others who speak (use) different languages and could feel a linguistic barrier as it is in real human communication.

As a means of overcoming the linguistic barrier, recently a couple of studies on the sign-language communication between avatars of different languages have been performed [2], [3]. In the communication, two avatars using different languages such as Japanese-Korean, Japanese-Portuguese, or Japanese-Chinese have communicated with each other using their own sign-languages. From the studies, it has been revealed that the gesture like sign-language can be used as an auxiliary communication means between different languages. However, the communication was done based on the arm-gesture only without considering the facial expression, even though it plays an important role in communicating messages [4].

In this paper, we propose an approach to the intelligent avatar communication based on the facial expression as well as the gesture animation. To produce the expression exactly on avatar mesh models, in the method, an adequate number of polygons are required. On the other hand, a smaller number of polygons allows for faster animation. In the avatar communication especially, the gesture and expression should be generated in real-time with the parameters (the joint angles and the action units). Moreover, avatars of various mesh models can participate in the communication. Considering these points, in this paper we employ three kinds of key-frame editors and develop a method of generating facial expression on various avatar models.

This paper is organized as follows: In Section 2, we briefly introduce the avatar communication between Korean and Japanese using the intelligent communication method. From Section 3, we discuss avatar models for the communication, three kinds of key-frame editors employed here, the movements of the cheeks and the jaws, physical constraints in the animation, and a method of generating emotional expression on different avatar models in succession. Experiments and discussions are provided in Section 7. Finally, the conclusions are given in Section 8.

2 Intelligent Avatar Communication between Different Languages

As mentioned previously, the sign-language communication between different languages has been considered as a means of overcoming the linguistic barrier [2], [3]. As an example, consider an avatar communication between Korean and

Japanese, where two avatars, named 'A' and 'B', communicate with each other in their own sign-languages. Korean Sing-Language (KSL) messages of 'A' are re- constructed (animated) with the corresponding Japanese Sign-Language (JSL) on 'B' after being transmitted as parameters of the joints angles and action units [3]. For example, the Korean avatar 'A' sends a KSL message, "NaNeun HakKyoEa KamNiDa (I go to school)", to 'B', then the Japanese avatar 'B' receives the corresponding message, which has been translated into JSL like "WatashiHa GakKouNi Ikimasu (I go to school)".

Both sign-language gestures of KSL and JSL are very similar to each other. The translation consists of two steps. First step is an analysis of the input sentence into intermediate sign-language words. The message is decomposed into three components: "Na (I)", "HakKyo (school)", and "GaDa (go)". Second step is a retrieval of sign-language parameters from KSL (or JSL) Dictionary, where the list of the sign-language words resulted from the previous step is utilized by keywords to find out sign-language parameters. The KSL (or JSL) Dictionary is composed of a set of Korean (or Japanese) Word Dictionary, a list of KSL (or JSL) animation parameters, and JP (or KR) pointers which are used as indexes for searching the corresponding JSL (or KSL) parameters from JSL (or KSL) Dictionary.

To achieve real-time communication, an intelligent communication method on a client-server architecture is employed. 3D avatar models are stored with the clients in advance and only intelligently coded data such as joint angles and action units are transmitted instead of motion pictures or their compressions. In the system, gesture and emotion images are analyzed into a sequence of parameters with the server and transmitted to clients through the Internet. Then, the corresponding images are reconstructed for clients with the received parameters on their models using CG animation techniques.

3 Avatar Models for the Communication

Sign-language gestures are composed of the arms' movements and the hands' (the fingers) shape. The movements are generated with changing the values of joint angles. Figure 1 shows two simplified coordinate systems for the left arm and the index finger, respectively. To simulate the arm's animation, we define the four joint angles (see the left picture of Figure 1), where θ_1 is the joint angle around the shoulder Z axis, γ is the joint angle around the shoulder X axis, θ_2 is the joint angle around the shoulder Y axis and θ_3 is the joint angle around the elbow Y axis.

We also define three joint angles for the hand's shapes (see the right picture of Figure 1), where θ_1 is the joint angle around the Z axis at the finger's MP (Metacarpal Phalangeal) joint, θ_2 is the joint angle around the Y axis at the MP (Metacarpal Phalangeal) joint, θ_3 is the joint angle around the Y axis at the PIP (Proximal Inter-Phalangeal) joint and β is the joint angle around the X axis at the DIP (Distal Inter-Phalangeal).

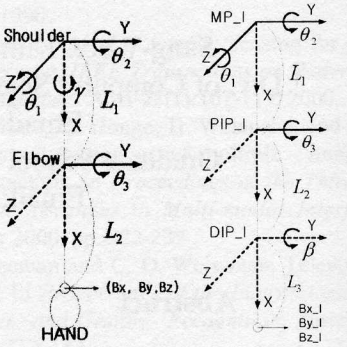

Figure 1: *Simplified coordinate systems for the left arm and the index finger. In the figure, the axes of dotted lines correspond to those of joint angles that don't need to be calculated. The details of the figure are found in the text.*

In the finger's coordinate system, from a geometric of the L_2, the L_3 and the joint angle β, a distance, L_2', from the PIP_I joint to the end of the finger, (B_x_I, B_y_I, B_z_I), is

$$L_2' = \sqrt{L_2^2 + L_3^2 - 2L_2L_3\cos(180 - \beta)}. \quad (1)$$

If we use the L_2' as the distance instead of the L_2 and L_3, in both coordinate systems, we can apply the same equations of inverse kinematics to obtain the joint angles of the arm and the hand (the finger). So, we can solve the problem of computing the fingers' joint angles in the same way as the arms' problem.

Then we employ some kinds of polygonal mesh head models, from which action units parameters are extracted and emotional expressions are generated. The facial components, such as the eyebrow, the upper eyelid, the eye, the lower eyelid, the nose, the upper and lower lips, and the jaw, are extracted in turn by using its polygonal data, geometric and color information. The larger the number of polygons increases, the better the quality of expression becomes. However, it takes more time to generate the animation.

4 Arm-gesture and Hand-shape Editors

The sign-language animation is generated by adjusting joint angles of the arms, the hands and the fingers. To generate the gesture animation, therefore, we first have to determine the values of these parameters. Based on the geometrical

analysis of the arm's or the finger's coordinate system, we calculate the values directly by using a transformation equation of inverse kinematics. In the inverse kinematics, given the position and orientation of an end-effect (the wrist or the end of the finger), we can obtain the values of joint angles of the arm (or the finger) so that the arm (or the finger) can be positioned as desired. Solution equations for the θ's are[1]

$$\theta_1 = \arctan\left(\frac{B_y}{B_x}\right), \quad (2)$$

$$\theta_2 = \arctan\left(\frac{S_2}{C_2}\right), \quad (3)$$

$$\theta_3 = \pi - \cos^{-1}\left(\frac{B_x^2 + B_y^2 + B_z^2 - L_1^2 - L_2^2}{-2L_1L_2}\right) \quad (4)$$

where L_1 and L_2 are the upper arm's length and the forearm's length, respectively. The S_2 and C_2 are constants that can be computed from the arm (or the finger)'s geometry [3]. When the position of an arm, (B_x, B_y, B_z), is given, the values of joint angles θ_1 , θ_2 and θ_3 can be calculated from Eqs. (2), (3) and (4), respectively. Then, the value the β can be decided by

$$\beta = c\theta_3, \quad (5)$$

where the c is an experimental constant. Finally the γ can be determined in heuristics.

Using the above equations, we design a kind of key-frame editor, by which the arm's gestures and the hand's shapes can be edited as desired. Figure 2 shows a view of the hand-shape editor, where the left is an editing window of gestures or shapes and the right is joint angles' panel. And the design window of the hand-shape editor can be changed into that of the arm-gesture editor by clicking a button. When we build up the desired key-frames by dragging a mouse on the left window, then we can acquire easily the corresponding parameter values (joint angle degrees) from the right panel.

The joints of arms and hands are coordinated by constraints that make some configurations impossible. To avoid unrealistic key-frames and reduce the search space, we apply the static and dynamic joint angle limits to the editors. In the human fingers, for example, it is nearly impossible to move the DIP joint without moving the adjacent PIP joint. Namely, a dependency between the joint angles of the DIP and PIP joints

$$\theta_{DIP} = \left(\frac{2}{3}\right)\theta_{PIP}, \quad (6)$$

can be utilized to avoid unrealistic motions.

Besides the dependency of Eq. (6), the movements of the MP joint such as the flexion, adduction and abduction are restricted by those of the neighboring fingers [6].

[1]The solution equations are derived from a 4×4 transformation matrix based on the arm (or the finger)'s geometry using inverse kinematics. A detailed derivation procedure can be found in [3].

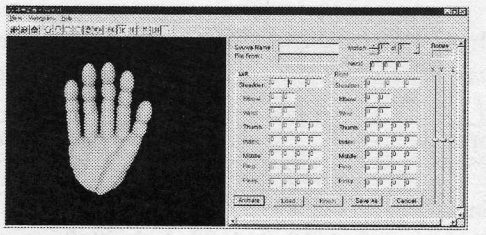

Figure 2: *View of a hand-shape editor employed here. The left window is shape builder and the right is their 20 (= 4 × 5) joint angles' panel. This editor can be changed into that of the arm-gesture by a click [5].*

5 Facial Emotion Editor

A widely used scheme for describing the facial expression is the Facial Action Coding System (FACS) [7], which describes the set of all possible basic facial muscle action units (AU's) performable by the human face. Figure 3 shows a facial emotion editor designed in this paper, from which principal action units of FACS for the universal expression can be extracted.

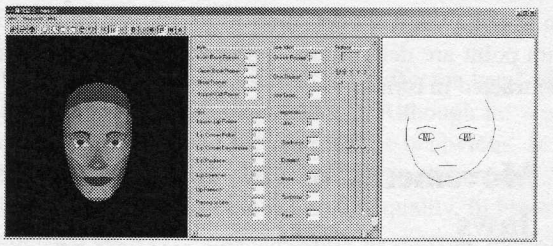

Figure 3: *View of a facial emotion editor. The left, right, and middle windows are, respectively, for 3D facial model, 2D comic-style model, and button panel for selecting an emotion and adjusting its intensity. The line-drawing 2D model is employed to assist the editing [8].*

With the editor, we can edit numerous expressions. For example, a "joy" expression can be edited with the combinations of AU's such as { AU1, AU2, AU12 }, { AU1, AU2, AU25 }, { AU1, AU2, AU12, AU25 }, etc. Among them, a few of the action units such as AU 1, AU 2 and AU 4, are used in common. On the other hand, each universal expression has one or two special AU's, which have the greatest influence on generating the expression. We define here the special action units as the *Principal Action Units (PAU's)* [8]. For example, the PAU of the "joy" expression is AU 12, by which we can significantly generate the expression. Table 1 shows a combination set of AU's for representing the six universal expressions and their PAU's.

In order to discriminate each expression easily, we also

Table 1: *A combination set of AU's for the six universal expressions, where the* $1+2+12$*, etc. represent the combinations of* { *AU1, AU2, AU12* }*, etc.*

Expression	Combination of AU's	PAU's
Joy	1+2+12, 1+2+25, 1+2+12+25	12
Sadness	1+4+15	15
Disgust	2+4+10, 2+4+25, 1+4+20+25	20
Anger	2+4+5, 2+4+15, 2+4+5+15	5+15
Surprise	1+2+5, 1+2+26, 1+2+5+26	26
Fear	1+4+20, 1+4+25, 1+4+20+25	48

define the *Expression Intensity* [8]. The intensity of a PAU corresponds with the amount of its activation and is defined as a value within $[0.0 \sim 1.0]$ according to its impressiveness. Higher intensity figures produce more angry (or joyful) expressions, which will be shown in the figures of the experimental results.

To generate an expression on a head model, first of all, we have to extract the corresponding PAU's from the model. To do this here, we utilize the three kinds of head model information: polygon information, coordinate information and color information. To extract AU and PAU parameters from a given mesh model, first of all, the top limit and the mouth point are determined. After that, facial components are extracted in turn.

6 Movements of the Cheeks and the Jaws

The movements of the cheeks and the jaws as well as the eyebrow, the eye and the mouth, play significant roles in generating the facial expression. In the previous studies, however, cheek and jaw movements have not been considered. So, we include the movements in generating the facial expression [8]. Some kinds of differential equations have been developed for different applications [9]. Among them, the parabolic partial differential equation and the Runge-Kutta numerical method can satisfactorily simulate the shape transformation of the cheeks and the jaws, respectively. To implement the movements, therefore, we utilize the two equations in this paper.

The parabolic partial differential equation for moving the cheek's mesh u_i is

$$u_i^{(j+1)} = u_i^{(j)} + \alpha \frac{\Delta t}{\Delta x^2}(u_{i+1}^{(j)} - 2u_i^{(j)} + u_{i-1}^{(j)}), \qquad (7)$$

where Δx and Δt are the mesh sizes of the x-direction and the t-direction, respectively, the i and j are the mesh point indices in the xt-plane and the α is a parameter which represents the height of the parabolic curve. The α, Δx and Δt are determined experimentally as values of $0.1 \leq \alpha \leq 10.0$, $\Delta x = D/(N-1)$ and $\Delta t < \Delta x^2/(2\alpha)$, respectively, where D is the diameter of the cheek's area and N is the mesh number of the x-direction. The left picture of Figure 4 shows a 3D view of the cheek's area to apply the parabolic partial differential equation.

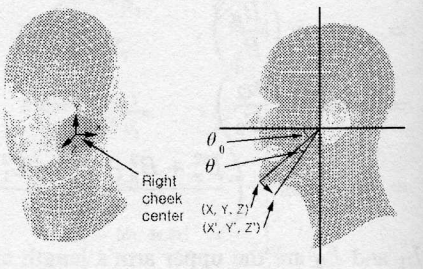

Figure 4: *The left picture is the extracted cheek's area and its coordinate axes, and the right is a geometry for computing the jaw's movements by using the Runge-Kutta method.*

In the right of Figure 4, the adjusted coordinate values, (X', Y', Z'), are

$$X' = X, \qquad (8)$$

$$Y' = d_n \times \cos(t_n), \qquad (9)$$

$$Z' = d_n \times \sin(t_n) \times \sin(\theta), \qquad (10)$$

where d_n and t_n are obtained by using the fourth order Runge-Kutta method with $d_0 (= y_0) = \sqrt{Y^2 + Z^2}$, $t_0 (= x_0) = \sin(\theta_0)$, the step size h and the number of steps N.

7 Experiments and Discussions

The arm-gesture editor, the hand-shape editor and the facial emotion editor, which have been implemented with the Visual C++ 5.0 and the Open Inventor on the Windows' platforms, were experimented. Using the arm-gesture editor, for example, the KSL message "NaNeun HakKyoEa KamNiDa (I go to school)" was edited into the five key-frames, f_1, f_2, f_3, f_4, f_5, from the beginning to the end. The parameter number of the arm-gesture editor for each key-frame, f_i, is eight. Since each parameter is an integer number (2 bytes long), the sub-total length of the parameters for the the five key-frames, $f_i, (i = 1, \cdots, 5)$, is amounted into 80 bytes $(= 2 \times 8 \times 5)$.

Using the hand-shape editor, sets of key-frames for numeric digits were produced. In this case, the number of extracted parameters for each key-frame, f_i, is forty (the 4 joint angle degrees per each finger, $4 \times 5 \times 2$ for two hands). By the same way as the case of the arm-gesture editor, the sub-total length of the parameters for the the five

key-frames, $f_i, (i = 1, \cdots, 5)$, is amounted into 400 bytes $(= 2 \times 40 \times 5)$.

The animation parameter transmitted is composed of parameters for facial expression as well as those of the arms and the hands. The facial expression parameters consist of the combination of principal action units and their intensities. The parameter number for an expression is the four or five (the three or four action units and a real number intensity, $2 \times 4 + 4 \times 1$). Therefore, the sub-total length of the facial parameters for the the five key-frames, $f_i, (i = 1, \cdots, 5)$, is amounted into 60 bytes $(= 12 \times 5)$. The total length of the arm-gesture, hand-shape and expression parameters to be transmitted is only amounted into 540 bytes $(= 80 + 400 + 60)$.

The message requires five key-frames and the frame number of in-betweens is 7. Since the delay time for one frame is 0.1, the amount of time to animate the message is 3.5 seconds $(= 5 \times 7 \times 0.1)$ and the transfer rate for the message is 1.23 Kbps $(= 540 \times 8/3.5)$. From these considerations, it is known that the amount of data sent is very small. The fact makes it possible to communicate in real-time between avatars using the proposed method.

A set of the key-frames were generated using the static and dynamic constraints as mentioned in Section 4. Figures 5 and 6 show the key-frames, where the left pictures are the key-frames generated without constraints and the right ones are edited with the constraints. A comparison shows that the editor equipped with the static and dynamic constraints can produce more natural key-frames than that of an editor without constraints.

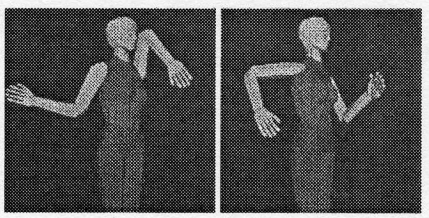

Figure 5: *Two key-frames generated by the arm-gesture editor. The left one is a key-frame edited without the constraints and the right is edited with the constraints. Among the two key-frames, the left is an impossible motion physically.*

Finally, the six kinds of universal expressions were reproduced with a 3,800 mesh head model. First, the facial components such as the eyebrows, the eyes, the mouth, the jaws and the cheeks are extracted in turns successively from a 3,800 polygonal mesh head model. Then, each expression was generated on them with the combination of AU's, PAU's and the intensities shown in Table 1.

Figures 7 and 8 show the "joy" expression reproduced on the head model without and with the movements of the cheeks and jaws as said in Section 6.

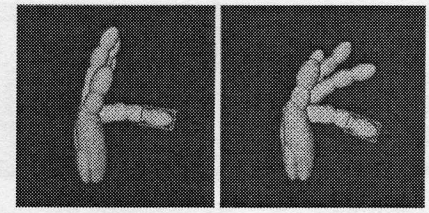

Figure 6: *Hand shape key-frames. The left of the pictures is a key-frame edited without constraints and the right is one edited with constraints. However, the left is an impossible motion physically.*

Figure 7: *The "joy" expression generated on the mesh model without the movements of the cheeks and the jaws, where their expression intensities are increasing from left to right.*

We repeated the experiments on different head mesh models: 1,900 and 296 mesh model. Although no significant difference between the expressions is obvious, in the figures, it is clear that the cheek and jaw movements method could be used to improve the expression quality. In Figure 7, we adjusted each PAU independently to generate the corresponding expression, which led to an unnatural expression. In Figure 8, however, we removed the artificiality by employing the parabolic partial differential equations and the fourth order Runge-Kutta method, where the movements of the cheeks and the jaws were adjusted dependently together with their surrounding regions.

8 Conclusions

In this paper we proposed an approach to the intelligent avatar communication based on the facial expression as well as the gesture animation. For the intelligent communication method, a method of extracting joint angle parameters and principal action units (PAU's) parameters was investigated. With the arm-gesture eitor employing the techniques of inverse kinematics and physical constraints, the joint angle parameters were founded efficiently. A small number of PAU's were extracted with the facial emotion editor using a comic facial model and partial differential equations.

From the results, we saw that the amount of data to be transmitted between avatars is very small and it is possible

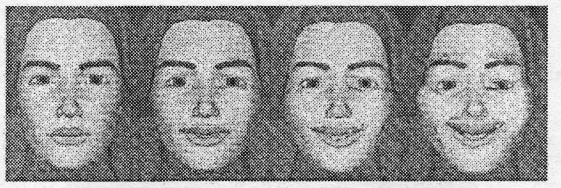

Figure 8: *The "joy" expression generated on the mesh model with the movements of the cheeks and the jaws, where their expression intensities are increasing from left to right.*

to communicate in real-time. Then, the quality of more realistic models was not as good when compared with the simple model even though the polygon number of the former was much larger than that of the latter. As an avatar model to be employed in cyberspace, therefore, a comic-style simple model may be better than the realistic complex models. In the comic model, the facial components of the eyebrow, the eyelid and the lips need a sufficient number of polygons while the other parts are described symbolically.

From the study, we confirmed a possibility that the proposed method could be useful for non-verbal communications between different languages in order to overcome the linguistic barrier in cyberspace. However, the current system is a preliminary system, which is used only in Korean and Japanese, not in English or any other languages. Research to extend the system, which have a lot of expressible words, are necessary to be carried out in the future.

Acknowledgments

This study was partially supported by a grant provided by the Ministry of Education, Culture, Sports, Science and Technology of Japan ("Grant-in-Aid for Scientific Research (A)(2)", no 13305026).

References

[1] B. Damer, *Avatars! Exploring and Building Virtual Worlds on the Internet* , Peachpit Press, 1998.

[2] Y. Aoki, R. Mitsumori, J. Li and A. Burger, "Sign language communication between Japanese- Korean and Japanese-Portuguese using CG animation," *Proceedings of ICASSP'98*, pp.3765–3768, May 1998.

[3] S.-W. Kim, J.-W. Lee and Y. Aoki, "Development of a sign-language communication system between Korean and Japanese using 3D animation techniques and intelligent communication method on the Internet," *IEICE Transactions*, vol. E83-A, no. 6, pp. 996–1004, Jun. 2000.

[4] A. Mehrabian, "Communication without words," in J. A. Devito, eds., *Communication: Concepts and Process*, Prentice-Hall, pp. 106–114, 1971.

[5] S.-W. Kim, Y.-W. Lee, J.-W. Lee and Y. Aoki, "A gesture-emotion keyframe editor for sign-language communication between avatars of Korean and Japanese on the Internet," *Proceedings of ITC-CSCC'2000*, Pusan, Korea, pp. 831 - 834, Jul. 2000.

[6] J. Lee and T. Kunii, "Model-based analysis of hand posture," *IEEE Computer Graphics and Application*, vol. 15, no. 5, pp. 77–86, Sept. 1995.

[7] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions*, vol. PAMI-22, no. 12, pp. 1424–1445, Dec. 2000.

[8] S.-W. Kim, J.-W. Lee, Y. Aoki and Y. Arakawa, "A comic expression method of universal emotions for intelligent avatar communications using principal action units," *Proceedings of ITC-CSCC'2001*, Tokushima, Japan, July 2001.

[9] E. Kreyszig, *Advanced Engineering Mathematics*, New York, John Wiley & Sons, Inc., pp. 1062–1110, 1988.